

AFRL-IF-RS-TR-2002-17
Final Technical Report
February 2002



HAWK: KNOWLEDGE ACQUISITION STARTING WITH NATURAL LANGUAGE

Massachusetts Institute of Technology

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. J774

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the U.S. Government.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

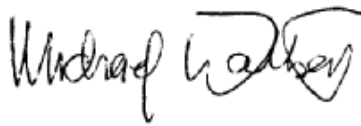
AFRL-IF-RS-TR-2002-17 has been reviewed and is approved for publication.

APPROVED:

A handwritten signature in black ink, appearing to read "John Spina".

JOHN SPINA
Project Engineer

FOR THE DIRECTOR:

A handwritten signature in black ink, appearing to read "Michael Talbert".

MICHAEL TALBERT, Maj., USAF, Technical Advisor
Information Technology Division
Information Directorate

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 074-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE FEBRUARY 2002	3. REPORT TYPE AND DATES COVERED Final Mar 98 - Dec 01	
4. TITLE AND SUBTITLE HAWK: KNOWLEDGE ACQUISITION STARTING WITH NATURAL LANGUAGE			5. FUNDING NUMBERS C - F30602-98-1-0036 PE - 62301E PR - IIST TA - 00 WU - 21	
6. AUTHOR(S) Boris Katz, Gary Borchardt, and Sue Felshin				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Massachusetts Institute of Technology 545 Technology Square Cambridge Massachusetts 02139			8. PERFORMING ORGANIZATION REPORT NUMBER N/A	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Defense Advanced Research Projects Agency 3701 North Fairfax Drive Arlington Virginia 22203-714			10. SPONSORING / MONITORING AGENCY REPORT NUMBER AFRL-IF-RS-TR-2002-17	
11. SUPPLEMENTARY NOTES Air Force Research Laboratory Project Engineer: John Spina/IFTD/(315) 330-4032				
12a. DISTRIBUTION / AVAILABILITY STATEMENT APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 Words) DARPA's Rapid Knowledge Formation program is developing new technology to automate the task of transforming raw human-understandable information into encoded, machine-understandable information. The project described in this report addresses a central subtask of this task: converting natural language text into and encoded representation that can support computer inference. The technical approach taken in this effort is based on two key insights: First, we can make the translation task manageable by breaking it into successive stages of isolating information, then standardizing it, then encoding it, with each stage facilitated by proven components of natural language processing technology. Second, we can gain leverage during the translation process by exploiting human interaction at a number of distinct points along the way.				
14. SUBJECT TERMS Natural Language Processing, Knowledge Bases, Knowledge Understanding			15. NUMBER OF PAGES 21	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UL	

Table of Contents

1. OVERVIEW	1
2. BACKGROUND	2
3. THE YEAR 1 START KNOWLEDGE BASE	4
4. THE BLITZ PREPROCESSOR.....	7
5. THE OMNIBASE DATABASE TOOL	8
6. THE YEAR 2 START KNOWLEDGE BASE	8
7. VIRTUAL COLLABORATION	12
8. NESTED ANNOTATIONS.....	13
9. DISTRIBUTED START PROCESSING.....	15
10. THE YEAR 3 START KNOWLEDGE BASE	15
REFERENCES.....	16

List of Figures

Figure 1:	Operation of the START system	4
Figure 2:	Sample START interaction	6
Figure 3:	Sample START interaction	6
Figure 4:	Sample START interaction	7
Figure 5:	Omnibase as integrated with START	9
Figure 6:	Sample START interaction	11
Figure 7:	Sample START interaction	11
Figure 8:	Sample START interaction	12
Figure 9:	Use of the virtual collaboration technique	13
Figure 10:	Nested annotations	14
Figure 11:	A START query that involves the use of nested annotations	14

1. Overview

This report describes work completed by the MIT Artificial Intelligence Laboratory in support of DARPA's High Performance Knowledge Bases program over the period from February 1998 to February 2001. The primary focus of the HPKB program is the design of tools for building large knowledge bases capable of supporting sophisticated question-answering and inference.

The project described in this report targets the construction of software tools that assist humans in accessing multi-media information from sources like the Internet, databases and knowledge bases. Such capabilities are critical to the notion of *human-centered problem solving*, in which computers collaboratively assist in the performance of interpretation, analysis and reasoning tasks by interacting largely on human terms and by responding to guidance when offered. In this project, special emphasis has been placed on enabling computers to interact on human terms, and thus a core software capability targeted by the project is that of efficiently accessing information on the basis of queries expressed in free natural language.

Supporting this effort is a key technology component grounded in sentence-level natural language processing. Whereas research in natural language processing has encountered significant difficulties in handling larger units of discourse, much progress has been made in mapping natural language phrases and sentences into sets of underlying semantic relationships that can be usefully manipulated by computers. Thus, this project takes the position of treating sentence-level natural language as itself a suitable representation for information content. This position is embodied in the notion of *natural language annotation*, whereby natural language phrases and sentences are used to organize and describe the content of arbitrary multi-media information segments, facilitating subsequent retrieval of those segments in appropriate circumstances when their annotations match human-submitted queries.

Sentence-level natural language processing, and specifically the technique of natural language annotation, are embodied in START, a software system capable of retrieving text, diagrams, images, information from databases, World Wide Web pages, and other types of information in response to natural language queries. Previous applications of START have included a system that answers questions of a geographical nature concerning several thousand cities and countries of the world, a system that answers questions about ongoing research at the MIT Artificial Intelligence Laboratory, and a system that answers questions about the U.S. mission in Bosnia-Herzegovina. START has been available for use through the World Wide Web since 1993 and has to date serviced over a million queries from users around the world.

During the HPKB program, START was applied in larger contexts, and new modules and functional capabilities were added to the system. A START knowledge base was constructed that contained information on military and economic capabilities and interests of various countries and organizations of the world, including oil production information, membership in international organizations, military assets and capabilities,

weapons strike capabilities, and various political, military and economic interests of countries, regional organizations and terrorist groups. Separately, two new modules were added to the START system. The first is a preprocessor, called Blitz, that quickly identifies names, places, numbers, times, dates, and other special classes of tokens. The second is a database tool called Omnibase that supplies START with data from internal database tables, data extracted from local text or HTML documents, data extracted from resources available through the Internet, and, more generally, data found by executing arbitrary segments of code. Finally, three functional capabilities added to the system include a notion of "virtual collaboration," in which START leverages native interfaces to available electronic resources; a notion of "nested annotations," in which annotations are placed within annotated material, enabling START to "look inside" multi-media information chunks to pull out contextually-presented information and compare this information across entities; and a mechanism whereby the START system is run as multiple, replicated processes that distribute the load of incoming queries among themselves. In the third year of the effort, an additional START knowledge base was constructed that contained information about banking and chemical concerns and individuals associated with those concerns. This domain serves as an analog to the domain of terrorist organizations and individuals associated with those organizations.

This report is divided into nine main sections, beginning with Section 2, which describes the START system. Following this, Section 3 describes the first year's START knowledge base construction effort. Section 4 describes the Blitz preprocessor that was incorporated into START during the first year's effort. Section 5 describes the Omnibase database tool that was incorporated into START during the first year's effort. Section 6 describes the second year's START knowledge base construction effort. Sections 7, 8 and 9 describe the notion of "virtual collaboration," the notion of "nested annotations," and the distribution of queries between replicated copies of the START system as implemented during the second year of the effort. Finally, section 10 describes the knowledge base construction effort conducted during the third year of the effort.

2. Background

START's information access capabilities result from a layering of capabilities. The simplest capability is that of indexing and retrieving English assertions. With this capability, the START system analyzes English text and produces a knowledge base which incorporates, in the form of embedded *ternary expressions*, the information found in the text. One can think of the resulting entry in the knowledge base as a "digested summary" of the syntactic structure of an English sentence. A user can retrieve the information stored in the knowledge base by querying it in English. The system will then produce an English response.

A representation mimicking the hierarchical organization of natural language syntax has one undesirable consequence: sentences differing in their surface syntax but close in meaning are not considered similar by the system. START solves the problem by deploying *S-rules* (in forward and backward modes) which make explicit the relationship between alternate realizations of the arguments of verbs.

The START system attempts to bridge the gap between current sentence-level text analysis capabilities and the full complexity of unrestricted natural language by employing *natural language annotations*. Annotations are computer-analyzable collections of natural language sentences and phrases that describe the contents of various information segments. START analyzes these annotations in the same fashion as any other sentences, but in addition to creating the required representational structures, the system also produces special pointers from these representational structures to the information segments summarized by the annotations. When START receives a question that matches an annotation, then, it can return as its answer the information segment associated with that annotation. This central capability of START is illustrated in Figure 1.

Suppose, for example, that a user wants to retrieve this text fragment related to the discovery of Neptune:

Neptune was discovered using mathematics. Before 1845, Uranus was widely believed to be the most distant planet. However, astronomers observed that Uranus was not always in the position predicted for it. The astronomers concluded that the gravitational attraction of a more distant planet was disturbing the orbit of Uranus. In 1845, John Adams, an English astronomer, calculated the location of this more distant planet. Urbain Leverrier, a French mathematician, independently did similar calculations. In 1846, John G. Galle and Heinrich d'Arrest of the Urania Observatory in Berlin, looked for the planet where Leverrier and Adams predicted it would be located. They saw the planet, which was later named Neptune, on September 23, 1846.

Let us assume that the sentence below serves as one of the annotations to this text fragment:

John Adams discovered Neptune using mathematics.

This means that START analyzed this sentence and incorporated it into the knowledge base along with a pointer to the text fragment. Now suppose the user asks one of the following questions:

Who discovered Neptune?

Did Adams discover Neptune?

How was Neptune discovered?

Was Neptune discovered using mathematics?

Tell me about Neptune's discovery.

START begins the process of answering any such question by creating a ternary expression to be matched against the knowledge base. It is important to emphasize that the full power of sentence-level natural language processing is brought to bear on the matching process. START's matcher works both on the *word* level (using, if appropriate,

additional lexical information about synonyms, hyponyms, IS-A trees, etc.) and on the *structure* level (utilizing necessary S-rules, information on verb-class membership, nominalization, etc.), although in the case of very simple questions such as those appearing above, most of this machinery is not utilized.

Since the representational structure returned by the matcher contains a special pointer to the annotated text fragment, START's usual sentence-level question-answering strategy is modified. Instead of passing the representational structure to the language generation, START simply follows the pointer and presents the text or multi-media fragment to the user.

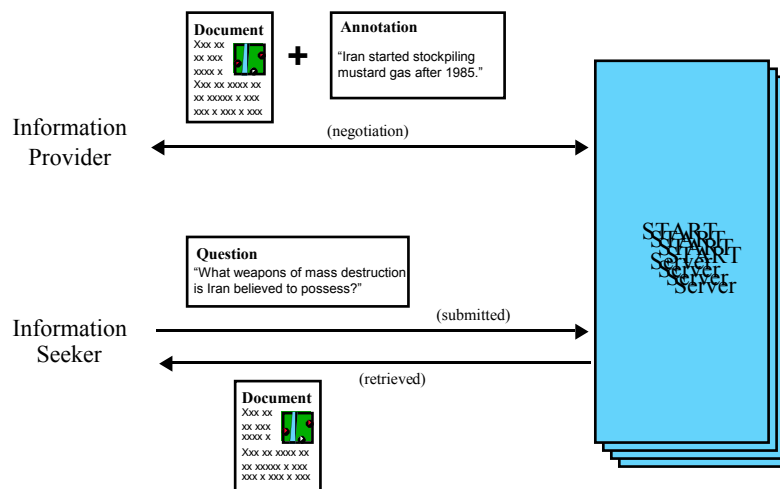


Figure 1. Operation of the START system.

3. The Year 1 START Knowledge Base

An important component of the HPKB program's first year evaluation effort was the Crisis Management Challenge Problem, in which participating knowledge-based systems were presented with the task of answering simplified-English questions concerning a fictitious scenario in which escalating hostilities between Iran and Saudi Arabia lead to Iran attempting to block all commercial shipping to and from the Persian Gulf. In support of this Challenge Problem evaluation task, a START knowledge base was constructed concerning the following five topic areas:

- 1997 CIA World Factbook information for 250 countries
- oil production information for Middle Eastern countries
- geographical coordinates, maps, distances and weather reports for several thousand cities and countries of the world
- membership and goals of several dozen international organizations
- military assets and capabilities of Persian Gulf countries

To enable this knowledge base to support natural language question answering, START's lexicon and set of S-rules were extended, and natural language annotations were attached to individual knowledge segments in the knowledge base. During the HPKB Crisis Management Challenge Problem evaluation itself, START answered 105 test questions posed to it. Following is a list of example questions answered by the START HPKB Server during the challenge problem testing period.

- Who are the members of Organization of Petroleum Exporting Countries?
- Can Iran attack oil ports bordering on the Persian Gulf with ballistic missiles?
- Which Unified Command of United States has responsibility for the Persian Gulf region?
- How many surface ships does the navy of the United States have in the Persian Gulf region?
- What amount of oil can be transported via pipeline from the Red Sea to the Mediterranean Sea?
- Can Kuwait attack targets in Iran with ballistic missiles?
- What cruise missiles does Iran have under development?
- How many missile craft does the navy of Saudi Arabia have?
- Which has the greater ratio between gross domestic product and population, Saudi Arabia or Iran?
- What are the agreed mutual member actions of the Organization of Arab Petroleum Exporting Countries?
- Does Saudi Arabia have a navy that can perform naval countermine warfare?
- Can Oman militarily threaten targets in the northern Persian Gulf region with ballistic missiles?
- What major oil refineries are in Libya?
- Of all countries in the Organization of Petroleum Exporting Countries, who has the greatest quota for production of oil?
- Is Iran a member of the Biological and Toxin Weapons Convention?
- What kinds of weapons of mass destruction is Libya believed to possess?
- Does Saudi Arabia have a deterrent against Iran's weapons of mass destruction?
- What is the Organization for the Prohibition of Chemical Weapons?
- How is the government of Iran different from the government of Saudi Arabia?
- Which peacekeeping forces of the United Nations have responsibility for Iraq?

Figures 2, 3 and 4 illustrate three question-answer instances from the START Year 1 knowledge base construction effort.

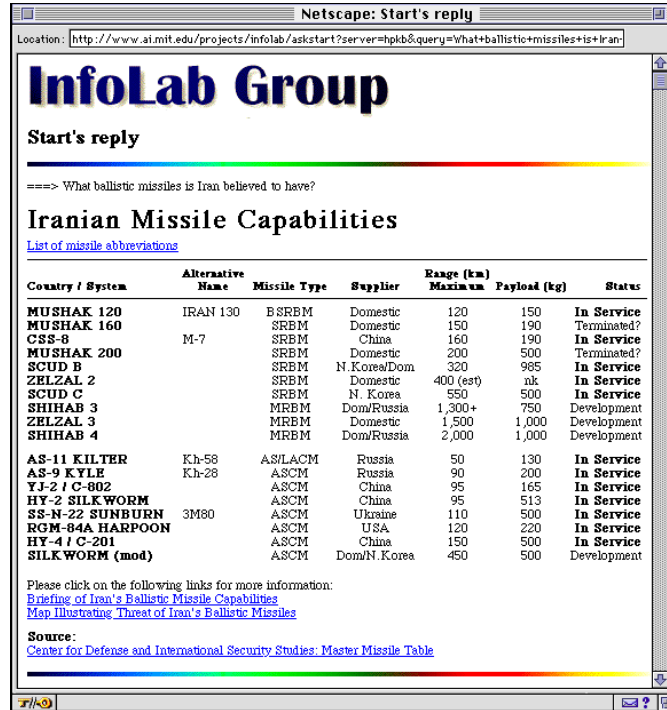


Figure 2. Sample START interaction.

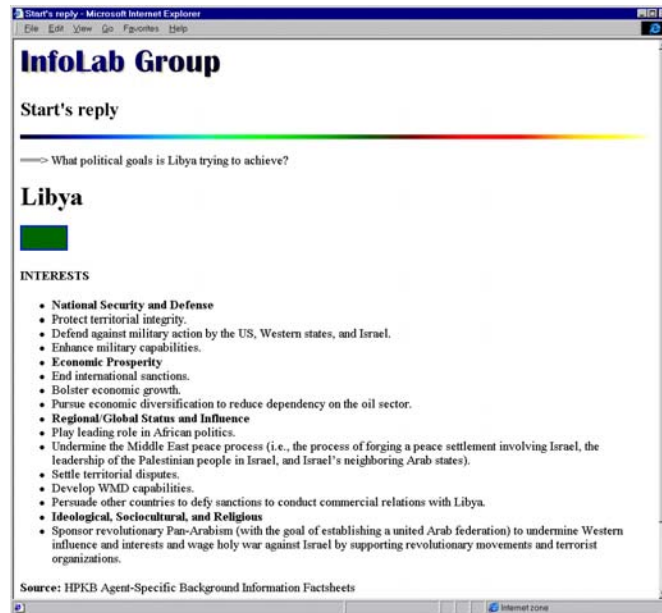


Figure 3. Sample START interaction.

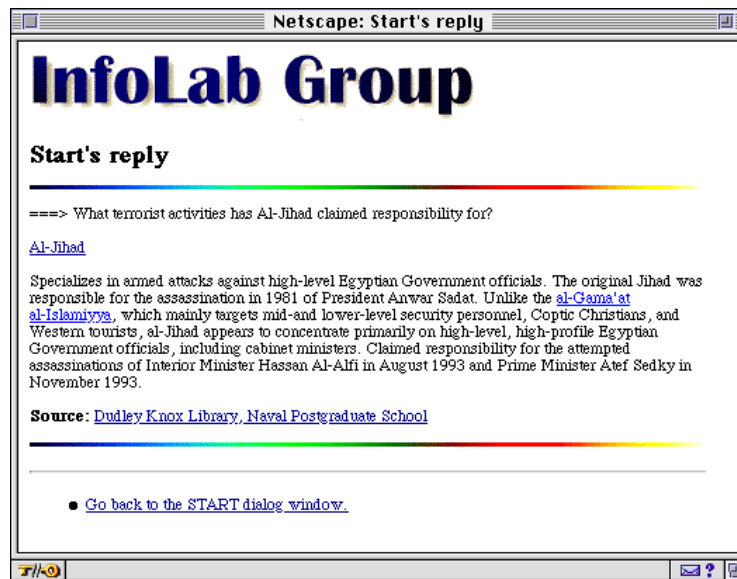


Figure 4. Sample START interaction.

4. The Blitz Preprocessor

A separate part of the START Year 1 HPKB effort was the integration of the Blitz syntactic preprocessor with the START system. Blitz acts as an independent, specialized token recognizer for common English fragments such as names, places, numbers, times and dates. Blitz tags these items with their assumed classification as a means of assisting START in its normal syntactic analysis. Blitz relies in part on the Omnibase system, described in the next section, to maintain tables of enumerated tokens like names and places. For tokens such as numbers, times and dates, Blitz uses internal heuristics to identify these items.

Following are several examples of output for the Blitz preprocessor regarding times, dates, addresses and quantities:

7:12 pm

(time "7:12 pm" :span (0 6) :hour 7 :minute 12 :time pm)

Friday, February 13, 1998

(date "Friday, February 13, 1998" :span (0 24) :day Friday :month February :date 13 :year 1998)

77 Massachusetts Avenue

(address "77 Massachusetts Avenue" :span (0 22) :number 77 :location "Massachusetts Avenue")

\$23.5 billion

(quantity "\$23.5 billion" :span (0 12) :value 2.35e+10 :unit \$)

two feet

(quantity "two feet" :span (0 8) :value 2 :unit feet)

5. The Omnibase Database Tool

Also during the Year 1 START HPKB effort, the Omnibase data server was integrated with the START system. Omnibase provides a uniform interface for processing queries of the form "Give me attribute X of object Y in class Z." The answers returned by Omnibase may be retrieved from internal database tables, extracted from local text or HTML documents, extracted from resources available through the Internet, or, in general, found by executing arbitrary segments of code. Omnibase uses two main procedures for storing data. First, it can add entities to its internal database. Entities are identified by their *class* and *symbol* name. Second, it can add scripts to its internal database for retrieving attributes of entities. Scripts are associated with classes.

Omnibase has two main procedures for retrieving data. First, given a query string, it can detect entities within the string, allowing mismatches in case and punctuation, and recognizing synonyms. Second, given an entity designator and an attribute name, it can retrieve the entity's value for that attribute by running the class script associated with the attribute.

Classes are specific to data sources, and therefore attribute scripts are customized to particular data sources. If one data source is a search URL, attribute scripts for classes in that data source will construct appropriate URLs, post them to the Web, and return the result, possibly parsing the result to return only part of it. If another data source is one or more Web pages with data for all entries expressed in similarly formatted HTML, attribute scripts for classes in that data source will retrieve the Web page and parse the HTML to find the correct segment of the Web page. The core of Omnibase is deliberately minimal in size. It is implemented in Guile and stores local data in an SQL database. Guile has all the advantages of Lisp and C combined, along with regular expression capabilities. Attribute scripts are extraordinarily easy to write, such that novice programmers can learn to write simple scripts in minutes.

Figure 5 illustrates the operation of the Omnibase system in combination with START.

6. The Year 2 START Knowledge Base

During Year 2 of the HPKB program, the Crisis Management Challenge Problem was expanded to include a larger set of information concerning interests of countries and other political entities, and, in particular, comparative information. In response, the START HPKB knowledge base was expanded to answer a broad range of new questions, as indicated by the following question grammar fragments:

Interests of international agents:

- "What are the <InternationalCategory> interests of <InternationalAgent>?"

Comparing interests of international agents:

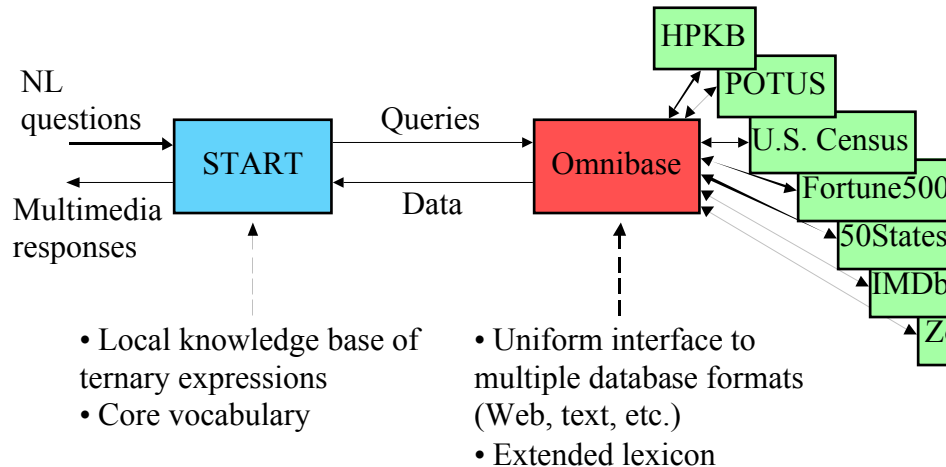


Figure 5. Omnibase as integrated with START.

- “How are the <InternationalCategory1> interests of <InternationalAgent1> {similar to, different from, related to} the <InternationalCategory2> interests of <InternationalAgent2>?”

where

<InternationalAgent> = countries..., terrorist groups..., criminal organizations..., regional organizations..., export control regimes..., nonproliferation regimes..., agent types...

<InternationalCategory> = {economic, trade, science, technology, military, diplomatic, political, cultural, social, religious, ideological, national security, defense, commercial}

Definitions of terms:

- “What is <Individual>?”

General comparisons:

- “How is <Individual1> {similar to, different from, related to} <Individual2>?”

where

<Individual> = international agents..., people..., places..., weapons...

In addition, START could answer a range of natural language variants of the questions indicated by these grammar fragments. During the HPKB Challenge Problem effort, START answered a total of 39 test questions submitted to the technology teams. The following are examples of some of these test questions:

- What is the Caspian Pipeline Consortium?
- Who is Richard Morningstar?
- How is Hamas like Hizballah?

- In the real world, what are the major components of economic interest of the UAE?
- How is Russia related to the UN?
- What potential motivations typically underlie a criminal organization's decision to smuggle military hardware and systems?
- In what major ways are Saudi Arabia's interests in the price of oil on the international market related to those of Iran?
- What are ways to deliver chemical weapons?
- What is the Azerbaijan International Operating Company?
- What is the Taliban?
- How is OPEC different from OAPEC?
- How is the GCC like OAPEC?
- What are subdivisions of the Gulf Cooperation Council?
- What relationship exists between PIJ and Israel?
- In the real world, what are the major components of economic interests for Iran?
- How is the Six plus Two group related to the United Nations?
- How is Russia related to the Caspian Pipeline Consortium?
- How is Azerbaijan related to the Caspian Pipeline Consortium?
- What are the defense interests of a typical criminal organization?
- What are the economic interests of a typical criminal organization?
- How are the typical interests of a terrorist group different from those of a criminal organization?
- How are the typical interests of a terrorist group like those of a criminal organization?
- In what ways are interests in deterrence against transnational terrorism for a country similar to interests in deterrence against proliferation of weapons of mass destruction for a country?
- In what major ways are Saudi Arabia's interests in relations with the United States different from those of Iran?
- In what major ways are Iran's interests in relations with the United States like those of Libya?
- What are ways to deliver biological weapons?

Figures 6, 7 and 8 illustrate three question-answer instances from the START Year 2 Crisis Management Challenge Problem evaluation exercise.

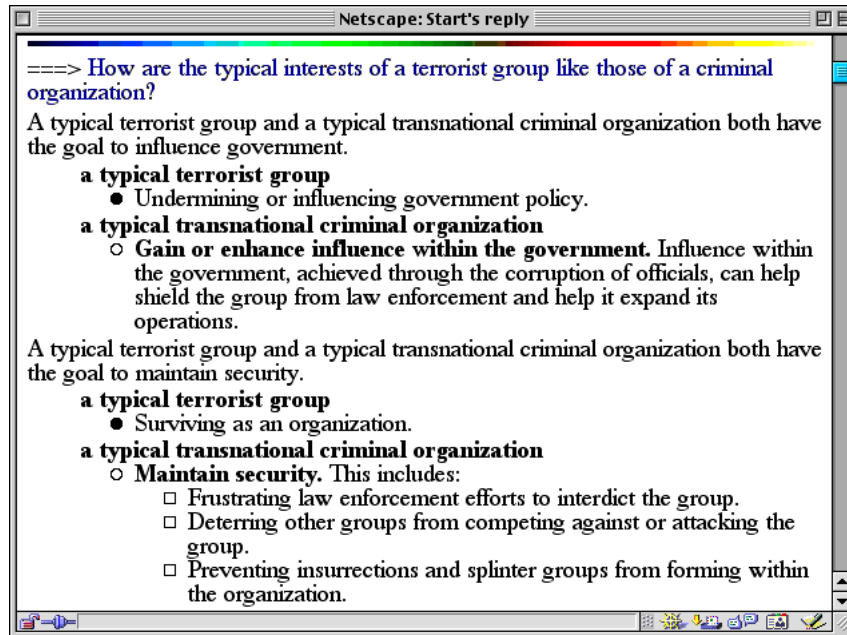


Figure 6. Sample START interaction.

====> What are ways to deliver biological weapons?

Biological Weapons

	<i>Nuclear</i>	<i>Biological</i>	<i>Chemical</i>
<i>Aerial Bombs</i>	X	X	X
<i>Artillery Shells</i>	X	X	X
<i>Ballistic Missiles</i>	X	X	X
<i>Cruise Missiles</i>	X	X	X
<i>Naval & Land Mines</i>	X		
<i>Torpedoes</i>	X		
<i>Unguided Rockets</i>	X	X	X
<i>Unorthodox Methods</i>	- Truck, Railcar, or Shipping Container - Man-portable "Backpack" Weapon	- Agricultural Sprayer or Atomizers - Simple Canisters, Bags, or other Containers	- Agricultural Sprayer or Atomizers - Simple Canisters, Bags, or other Containers

Table 2: WMD Agent Means of Delivery

Figure 7. Sample START interaction.

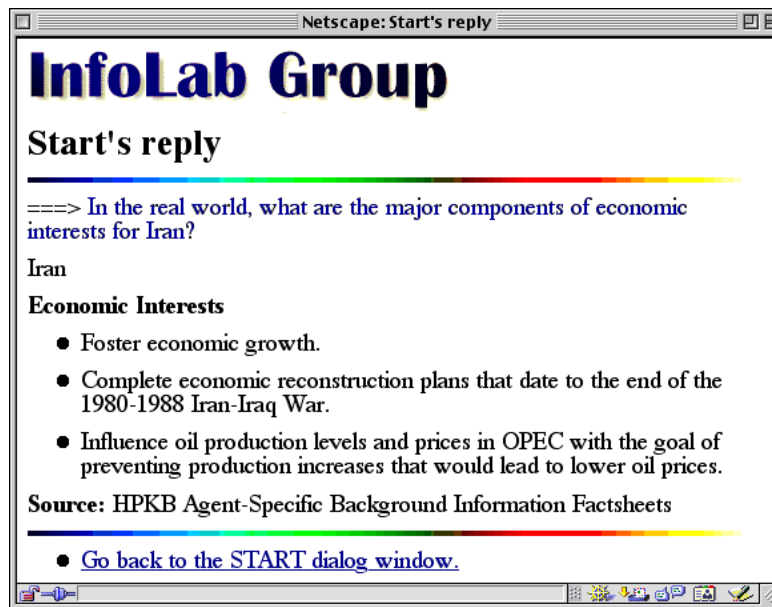


Figure 8. Sample START interaction.

7. Virtual Collaboration

A sizeable portion of the START research effort concerns the development of efficient procedures by which humans and computers may collaboratively place new information in service for language-based access. An especially important technique is that of *virtual collaboration*, in which we take advantage of highly-structured electronic information resources by implementing small program fragments that translate natural language queries to resource-specific interface queries. In this manner, START can leverage the work of others in a “virtual” manner, using the fruits of labor of a large group of people without explicitly collaborating with them.

Use of the virtual collaboration technique is highly efficient: in many cases, hundreds or thousands of new questions are made answerable through a minimal amount of additional human effort. This integrates with START’s existing technological base in a logical manner, as follows. For highly-structured information sets such as databases, tables, and form-based Web resources, virtual collaboration is employed where possible. For less-highly structured information resources such as conventional multi-media documents, individual natural language annotations are attached directly to components of information that we wish to make accessible. Finally, for largely-unstructured information sets such as lists of unrelated facts, individual natural language assertions are stored in START’s knowledge base for subsequent retrieval.

The Omnibase system plays an important role in START’s use of the virtual collaboration technique. First, Omnibase provides uniform access to external information sources, be they text documents, web pages, databases, or other types of

resources. Specially-constructed access scripts translate Omnibase queries into resource-specific queries or answer-determination procedures. Second, Omnibase makes tokens in external databases available to START as if they were part of START's native vocabulary.

Figure 9 illustrates a simple START query that makes use of the virtual collaboration technique to access information on the Internet Movie Database Web site.

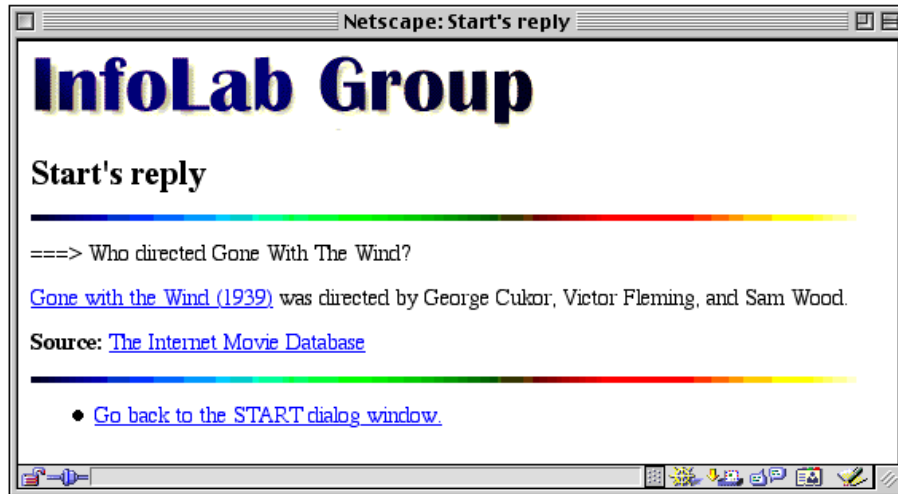


Figure 9. Use of the virtual collaboration technique.

8. Nested Annotations

The Year 2 Crisis Management Challenge Problem included a number of question types that required a comparison of different quantities. As part of START's solution to this problem, a new technique called "nested annotations" was developed. This technique builds on START's standard information access approach of attaching content-describing phrases and sentences (annotations) to multi-media information chunks. Nested annotation involves a hierarchical insertion of annotations within annotated material. START can then use the contained annotations to "look inside" retrieved annotated material in order to compare multi-media information chunks to one another.

Figure 10 illustrates the use of the nested annotation technique. When START is presented with the query "How are Iran's political interests similar to those of Libya?", it first retrieves information segments that have annotations indicating that they are descriptions of Iran's political interests and Libya's political interests. Next, START looks within the retrieved information segments to a second set of annotations that summarize particular interests of those countries. By matching the specific-interest-specifying annotations to one another, START can determine which interests are held in common between the two countries. The annotated material itself can then be inserted

into the returned answers as elaborations of the similar and dissimilar interests determined by START. Figure 11 illustrates the answer returned by START when posed with the above question.

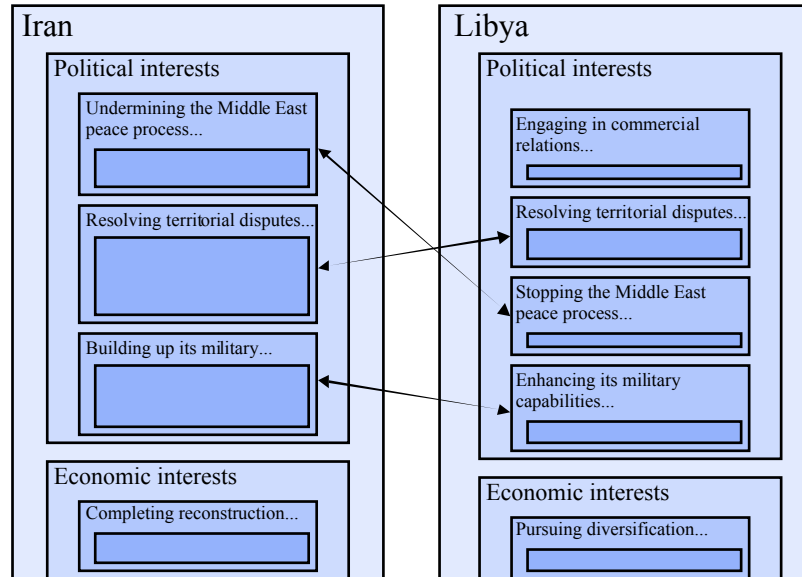


Figure 10. Nested annotations.

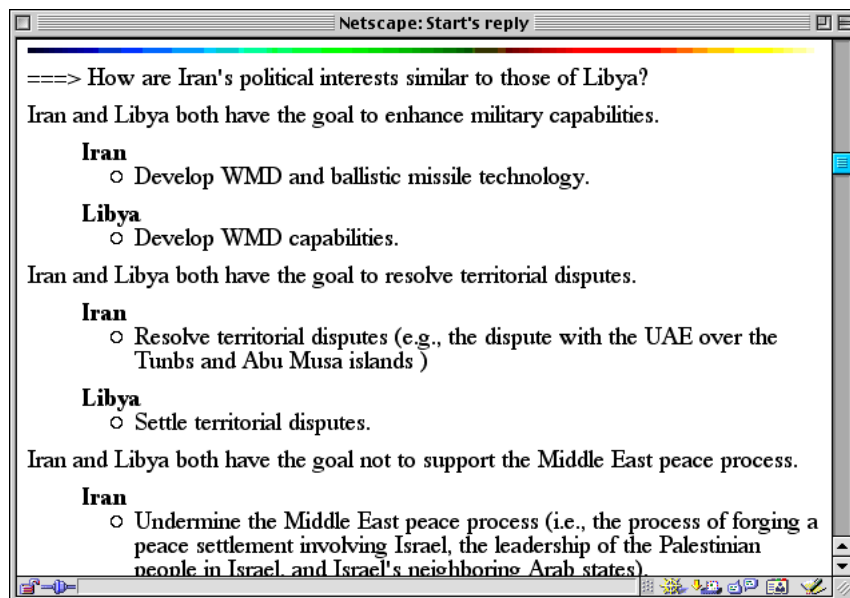


Figure 11. A START query that involves the use of nested annotations.

9. Distributed START Processing

An additional capability realized within the START system during Year 2 of the START HPKB effort concerns the replication of START processes for handling incoming queries. By this technique, a number of START and Omnibase servers collectively distribute the load of queries submitted to START over the World Wide Web. As a result, if particular queries take longer to process---for example, queries that generate requests to some external resources available on the World Wide Web---then only those queries are held up, while other users' queries are processed independently.

10. The Year 3 START Knowledge Base

In Year 3 of the START HPKB effort, a distinct knowledge base was constructed in support of research activity proceeding under an initial "Jumpstart" effort for DARPA's Evidence Extraction and Link Discovery (EELD) program. This program concerns the identification of important facts and patterns from large collections of text and structured information sources, as is particularly relevant to the detection and tracking of terrorist activities. During the Jumpstart effort, a base of information related to banking and chemical concerns was used as a stand-in for actual information on terrorist organizations and their activities. START was made to provide access to a substantial portion of this base of information plus information about the EELD effort itself, resulting in START's ability to answer a range of relevant questions such as the following:

- Tell me about ABC Bancorp.
- Who is the president of ABC Bancorp?
- Who is the CFO of Alliance Bancorp?
- Who are the subsidiary officers of ANB Corp?
- Who was the president of Hinsdale on May 5th 1994?
- Does Bristol sit on the board of directors at ABC Bancorp?
- What is the address of Banknorth Group Inc.?
- How can I contact Bancorp Connecticut Inc by phone?
- Where does Fredric G. Novy work?
- What is Michael Weeks' position at BancorpSouth Inc?
- Who are the directors of Kenneth Hunnicutt's company?
- What is the stock ticker of Kenneth Hunnicutt's company?
- After Cragin was acquired by ABN-AMRO North America, where did Novy go?
- Does Novy own any stocks?
- How many shares of Alliance does Fredric Novy own?
- How much money does Albert Eckert earn at First Bell Bancorp?
- How much money does Bristol make?
- How many directors does Alliance Bancorp have?
- What is the ticker for ANB Corp?
- How many employees does ANB Corp have?
- What is Alliance's Primary SIC code?
- What is the current status of Alliance Bancorp?

- What happened to Lionel Tokioka?
- Who got nominated for director at Alliance?
- How many presidents has Liberty Bancorp had in the last 5 years?
- Who owns more than 10% of the stocks of Liberty Bancorp?
- Who are the insiders at Alliance Bancorp?
- Give me a list of Liberty Bank's subsidiaries.
- Why did Hinsdale merge?
- What is Hinsdale's new name after merging with Liberty Bancorp?
- What do you know about banks?
- Do you have any information about the WTC bombing?
- How many banking scenarios are there?
- What do you know about Alliance Bancorp?
- What data is available for information extraction researchers?
- What handouts were distributed at January's meeting?
- Who is working on SHIELD?
- What is Gunning's email?
- How can I get in touch with Chris White?
- How do I contact Tom?
- What groups is Rosie Jones a member of?
- What university is Daphne Koller at?
- Show me the SHIELD slides.

As part of the Year 3 START knowledge base construction effort, START was supplied with natural language annotations, lexicon entries and syntactic/semantic transformation rules, enabling it to answer each targeted question type in a number of variant forms. The resultant START question-answering system was then installed on the World Wide Web, making it accessible to interested users.

References

The following articles were published during the course of the START HPKB effort.

Boris Katz, Deniz Yuret, Jimmy Lin, Sue Felshin, Rebecca Schulman and Adnan Ilik, "Blitz: A Preprocessor for Detecting Context-Independent Linguistic Structures," Proc. Pacific Rim International Conference on Artificial Intelligence (PRICAI '98), Singapore, November, 1998, pp. 13-18.

B. Katz, D. Yuret, J. Lin, S. Felshin, R. Schulman, A. Ilik, A. Ibrahim, and P. Osafo-Kwaako, "Integrating Web Resources and Lexicons into a Natural Language Query System," Proc. Sixth IEEE International Conference on Multimedia Computing and Systems, Florence, Italy, June 1999.

B. Starr, V. Chaudhri and B. Katz, "HIKE – A Query Interface and Integrated Knowledge Environment for HPKB," Proceedings AAAI 16th National Conference on Artificial Intelligence, Orlando, FL, 1999, p. 985.

O. Uzuner, B. Katz and D. Yuret, “Word Sense Disambiguation for Information Retrieval,” Proceedings AAAI 16th National Conference on Artificial Intelligence, Orlando, FL, 1999, p. 927.

B. Katz and J. Lin, “REXTOR: A System for Generating Relations from Natural Language,” in Proceedings ACL 2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval, Hong Kong University of Science and Technology, October 2000.